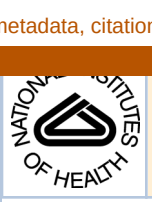

Published in final edited form as:

Environ Sci Technol. 2012 March 6; 46(5): 2772–2780. doi:10.1021/es203152a.

Integrating Address Geocoding, Land Use Regression, and Spatiotemporal Geostatistical Estimation for Groundwater Tetrachloroethylene

Kyle P. Messier, **Yasuyuki Akita**, and **Marc L. Serre***

Department of Environmental Science and Engineering, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC

Abstract

Geographic Information Systems (GIS) based techniques are cost-effective and efficient methods used by state agencies and epidemiology researchers for estimating concentration and exposure. However, budget limitations have made statewide assessments of contamination difficult, especially in groundwater media. Many studies have implemented address geocoding, land use regression, and geostatistics independently, but this is the first to examine the benefits of integrating these GIS techniques to address the need of statewide exposure assessments. A novel framework for concentration exposure is introduced that integrates address geocoding, land use regression (LUR), below detect data modeling, and Bayesian Maximum Entropy (BME). A LUR model was developed for Tetrachloroethylene that accounts for point sources and flow direction. We then integrate the LUR model into the BME method as a mean trend while also modeling below detects data as a truncated Gaussian probability distribution function. We increase available PCE data 4.7 times from previously available databases through multistage geocoding. The LUR model shows significant influence of dry cleaners at short ranges. The integration of the LUR model as mean trend in BME results in a 7.5% decrease in cross validation mean square error compared to BME with a constant mean trend.

Introduction

Geographic Information Systems (GIS) based techniques are well known, cost-effective, and efficient methods implemented in exposure assessment and epidemiologic studies.[1–3] Applications include address geocoding, land use regression (LUR), and geostatistics, which can be utilized for defining at-risk populations,[2] identifying explanatory variables,[4–7] and interpolating concentration and exposure values to unmonitored locations,[8,9] respectively. Many researchers, including academia and state or national government level regulators, have utilized one of these methods at any given step in their studies. For instance, a study of air pollution and lung cancer in Stockholm, Sweden[10] used GIS techniques such as address geocoding to facilitate exposure estimations. Su et al.[7] demonstrated a successful approach for statistically modeling NO_2 exposure using LUR techniques. Similarly LUR has been used to identify sources of contamination in groundwater.[4,5,6,11] The advanced non-linear, non-Gaussian Geostatistical framework known as Bayesian Maximum Entropy (BME) has

*Corresponding Author: Marc L. Serre Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina, 1303 Michael Hooker Research Center, Chapel Hill, NC 27599 Phone: (919) 966-7014 marc_serre@unc.edu Fax: (919) 966-7911.

Supporting Information Available Details on data locations, estimating below detects, and the covariance model is provided in the Supporting Information. Supporting information is available free of charge via the Internet at http://pubs.acs.org/

proven successful in the space/time estimation of surface water[8] and air quality[9] concentrations at unmonitored locations.

These studies implement the GIS techniques in a sound manner; however, there is still the need to improve concentration estimates for epidemiologic research. We hypothesize that concentration estimates can be significantly improved by integrating multiple GIS based techniques. For example, address geocoding can be used in conjunction with LUR or geostatistics to increase the available data to the models. Similarly, geostatistical theory allows for the incorporation of mean trends that can be based on other model output, thereby allowing LUR to be incorporated as a physically meaningful mean trend. Furthermore, LUR techniques introduced by Su et al.[7] that help determine explanatory variable ranges can be incorporated into an unbiased non-linear geostatistical estimator such as BME.

In our case example, we address a critical mission of the North Carolina Division of Water Quality (DWQ), which is to continually assess the quality of the state groundwater in order to ensure that the state has a clean water supply. However, budget limitations prevent the DWQ from maintaining more than 300 ambient ground water monitoring wells across the state. For contaminants with localized plumes 300 monitoring wells and one GIS based technique (LUR or BME) are not sufficient for accurately estimating ground water contamination across North Carolina.

To address this common shortfall this paper demonstrates an integration of GIS based techniques: address geocoding, LUR, and the BME framework. We utilize address geocoding to increase available ground water data; land use regression to identify significant contaminant sources and hydrogeological flow; and BME to increase concentration estimation accuracy and better preserve rank-order in estimations.

We show our multi-stage framework with groundwater Tetrachloroethylene (PCE) data across North Carolina. PCE is a chlorinated solvent that is commonly used for dry cleaning fabrics and for metal degreasing[12], and “likely carcinogenic to humans” according to the United States Environmental Protection Agency (USEPA).[13] PCE is an ideal candidate for further concentration analysis because the available monitoring data is sparse and epidemiology studies are not conclusive in regards to its carcinogenicity.[13] The proposed framework utilizes the strengths of the three aforementioned methods leading to more accurate concentration estimates for epidemiology studies, which to the authors' knowledge, has not been shown in any exposure analysis in the United States. We estimate concentration, which serves as a proxy for exposure since we are dealing with untreated groundwater. Lastly, we conclude about the relevance of this work on epidemiology studies and groundwater PCE exposure.

Materials and Methods

Tetrachloroethylene Data Sources

Data on groundwater PCE were compiled from three sources, which are detailed as follows:

North Carolina monitors PCE through the Dry Cleaning and Solvent Cleanup Act (DSCA) section of the N.C. Division of Waste Management, which was established to help fund cleanup of PCE contamination.[14] DSCA maintains contracts with private companies to construct monitoring wells, which in turn provide DSCA with an electronic data deliverable (EDD) that contains the locations of PCE concentrations in monitoring wells. There are approximately 207 DSCA sites distributed across the state, but EDD's are not available for all the sites yet. For this study, we have data from 96 DSCA monitoring sites, collected from 1999–2010, resulting in 1062 monitoring wells with 2356 space/time samples. It should be

noted that the DSCA monitoring sites are spatially clustered since all of the monitoring wells are located around known polluted sites.

The North Carolina Department of Health and Human Services collects volatile organic carbon data from North Carolina homeowners. Prior to 2007 the data collected were from homeowners who voluntarily had their well tested. Starting in 2007 all new wells built were required by law to be tested.[15] The data are analyzed at the Department of Public Health State Lab, where a paper report for each well is created and stored. There is no standard for providing GPS coordinates in the report; however, the well address is provided. Consequently, we digitized the paper reports and then applied a geocoding scheme to obtain geographic coordinates. Using the address locator tool of ArcGIS 9.3 (ESRI, Redlands, CA), data were assigned coordinates in a multi-stage process using a North Carolina point reference file (courtesy of DHHS spatial analyst group), followed by a North Carolina Department of Transportation line reference, then with a U.S. street address line reference file. All geocoded addresses with a match score of 70 and above were included in the dataset. The address geocoding resulted in 2,411 geocoded wells with 2,874 space/time samples (out of 4,102) from the years 2003–2010 that were previously unavailable.

We downloaded all of the PCE well data available from the USGS NWIS website.[16] We obtained 71 monitoring wells with 94 space/time samples from 2001–2010 distributed across the state. Our blending of data sources resulted in 4,119 unique wells with 5,402 space/time samples (Figure S1). Concentration statistics for each source is also provided in supporting information (Table S1)

Land Use Regression Model

Modeling a contaminant source global mean trend in this framework serves four main purposes: (1) To identify point sources that significantly affect groundwater PCE, (2) to investigate the range of influence of point sources on the dependent variable,(3) To account for hydrogeological flow direction, and (4) to provide the geostatistical model with a well-informed mean trend, as opposed to common techniques such as a constant mean trend. We present the details of a LUR for our case of PCE in North Carolina; however, the principles are applicable for other contaminants and other types of statistically based explanatory variables.

We model the global mean trend of groundwater PCE using a LUR model, where the dependent variable is the log-transformed PCE concentration. By taking the log-transformation we reduce the skewness from 7.07 to 0.78.

Our PCE monitoring data contained below detect data; therefore a method to account for samples without detectable PCE was necessary. There are a variety of acceptable methods to handle environmental data containing below detects (i.e. left-censored data), including assigning the below detect a value of half the detection limit[8] or performing the analysis based on detection frequency.[11] These methods, however, only incorporate information about the detection limit, and ignore valuable information provided by the above detect measurements. In this study we introduce a novel two stage approach: First we characterize the *population* distribution of log-PCE based on above detects. We model the probability distribution function (PDF) of log-PCE using a Gaussian distribution with a mean μ and variance σ^2 such that the cumulative distribution function (CDF) at the detection limit and the 95th percentile produce values equal to the percent of samples below detect and the 95th percentile of the sampled values, respectively. A full numerical description for the technique is described in the Supporting Information. This PDF characterizes the *population* distribution of all log-PCE concentrations (dashed line, Figure 1). Then for each below detect measurement, we know that log-PCE is a value drawn from the population

distribution truncated above the detection limit. Hence we assign to each below detect a value equal to the mean of the Gaussian distribution truncated above the detection limit (plain line, Figure 1).

PCE almost always occurs because of anthropogenic causes,[12] thus we constructed the independent variable based on the locations of sites that are known or potential sources of PCE.

The location and associated information for land use variables were obtained from NC Division of Waste Management GIS personnel and from NC Onemap,[17] a public online database for GIS data. We incorporate the following land use variables into our contaminant source database: dry cleaners including DSCA and non-DSCA sites; Resource Conservation and Recovery Act (RCRA) hazardous waste generator sites; RCRA sites with known releases of PCE according to the EPA Toxic Release Inventory; Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA or Superfund) sites; National Priority Listing sites with known contamination of PCE; National Pollutant Discharge Elimination System (NPDES) sites; septage land application sites/ septage detention or treatment facility sites (Septage); brownfield sites; landfills (current and pre-regulatory); and manufacturing gas plants (MGP) sites.

It is generally believed that major types of sources of PCE include dry cleaners, hazardous waste generators and Superfund sites, but other types of sources cannot be discounted.[12] For each type of pollution source l, (e.g. l=dry cleaners) we construct an explanatory variable calculated as the cumulative exponentially decaying contribution from each polluted site of that type, which can be expressed as

$$X_i^{(l)} = \sum_{j=1}^{n} C_{0j} \exp\left(-3 * \frac{D_{ij}}{a_j}\right) \quad (1)$$

where $X_i^{(l)}$ is the contamination contribution at well i from source l, C_{0j} is the initial concentration at the polluted site j, D_{ij} is the distance between well i and polluted site j, n is the total number of polluted sites of type l, and a_l is the exponential decay range defining the range of influence of that type of pollution source. The exponential operator in the model ensures concentration decreases quickly as the distance increases from the contaminant source. The cumulative aspect of the model accounts for the density of contaminant sources. We have information on contamination for dry cleaners, thus for dry cleaners C_{0j} is the maximum concentration sampled at site j. Since no information is known about concentrations for any other pollution sources then C_{0j} is assumed constant across all sites other than dry cleaners.

The dependency of groundwater PCE log-concentration, Z, with different types of known sources can be expressed for sample i as

$$Z_i = \beta_0 + \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \cdots + \beta_m X_i^{(m)} + \epsilon_i \quad (2)$$

where Z_i is the log-PCE concentration for sample i, $X_i^{(1)}$ through $X_i^{(m)}$ are explanatory variables representing different types of contaminant sources $\beta_0, \ldots, \beta_m$ are linear regression coefficients, and ϵ_i is an error term. This model allows investigation into the effects of various types of contaminant sources as well as the range of influence, a_l, associated with each type of source. First, we investigate the effects of each a_l individually by constructing a series of univariate models for each pollution type l, and exploring how the univariate coefficient of determination r^2 changes as a function of each decay range a_l.

Then we use step-wise regression to determine if multivariate models increase the model accuracy. We choose the regression model with maximum r^2 obtained with physically meaningful (i.e. positively valued) and statistically significant regression coefficients.

Note that the model described thus far does not account for hydrogeological flow. Information about well depth is not available for all the monitoring data, but we can use an elevation raster grid to account for groundwater flow and transport using a flow accumulation algorithm.[18] The exponentially decaying model (Eq. 1) is used to calculate the input PCE concentration used for the flow accumulation. The flow accumulation algorithm is used to calculate the contaminant concentration resulting from hydrogeological transport of the contaminant along flow lines following the downward elevation gradient. We use flow accumulation with multiple direction flow routing[18], and with a runoff ratio[18] proportional to the elevation gradient, so that steeper gradients will lead to transport of contaminant over longer distances than that on flat gradients. The corresponding estimated regression coefficients $\widehat{\beta}_1, \ldots, \widehat{\beta}_m$ obtained for that model can then be used to construct the LUR model $L_Z(s)$ of log-PCE concentration at any spatial location $s=(s_1,s_2)$ as

$$L_Z(s) = \widehat{\beta}_0 + \widehat{\beta}_1 X^{(1)}(s) + \widehat{\beta}_2 X^{(2)}(s) + \cdots + \widehat{\beta}_m X^{(m)}(s), \quad (3)$$

where $X^{(1)}(s), \ldots, X^{(m)}(s)$ are the flow routed cumulative exponentially decaying contribution from each type of pollution sources calculated for the spatial location s. Once the LUR model $L_Z(s)$ is calculated using β coefficients obtained with the ordinary least squared estimator, we account for the auto-correlation in the LUR estimate errors by recalculating $L_Z(s)$ using β coefficients recalculated from a generalized least squared estimator.

BME Estimation Framework for Space/Time Mapping Analysis

In this study we use the BME method of modern spatiotemporal geostatistics[19–21] to estimate the concentration of groundwater PCE across space and time. *BMElib*,[22,23] a powerful MATLAB numerical toolbox of modern spatiotemporal geostatistics implementing the BME theory, was used to create space/time maps of PCE concentration across North Carolina. BME is a space/time geostatistical estimation framework grounded in epistemic principles that reduces to the space/time simple, ordinary, and universal Kriging methods as its linear limiting case when considering a limited, Gaussian, knowledge base, while also allowing the flexibility to process a wide variety of additional knowledge bases (physical laws, empirical relationships, non-Gaussian distributions, hard and soft data, etc.) that are beyond the reach of the Kriging methods of linear geostatistics.[8,9,21] We only provide the fundamental BME equations for mapping PCE; the reader is referred to other works for more detailed derivations of these equations.[18,19,23]

The theory of space/time random field (S/TRF) is used to model the variability and uncertainty associated with the distribution of PCE concentration across space and time. Our notation for variables will consist of denoting a single random variable Z in capital letter, its realization, z, in lower case; and vectors and matrices in bold faces, e.g. $\boldsymbol{Z} = [Z_1, \ldots, Z_n]^T$ and $\boldsymbol{z} = [z_1, \ldots, z_n]^T$. Let $Y(p)$ be the S/TRF describing the distribution of PCE concentration across space and time, and let $Z(p) = \log Y(p)$ be its log-transform, where $p = (s,t)$, is the space coordinate and t is time. The log-transformed residual S/TRF is defined as

$$X(p) = Z(p) - m_z(s) \quad (4)$$

where $m_Z(s)$ is a global geographical trend. In this work, we first use a constant global geographical trend, and we then compare that approach with using $m_Z(s) = L_Z(s)$, the integrated land use model.

The knowledge available is organized in the general knowledge base (G-KB) about the S/TRF $X(p)$ (e.g. describing its space/time variability, mean, covariance, etc.) and the site-specific knowledge base (S-KB) corresponding to the hard and soft data available at a set of specific space/time points p_d.

The G-KB for the S/TRF X_d describes its local space/time trends and dependencies. In this work, the general knowledge consists of the space/time mean trend function $m_x(p) = E[X(p]$, and the covariance function $C_x(p,p') = E[[X(p - m_x(p)][X(p') - m_x(p')]]$ of the S/TRF $x(p)$.

A key conceptual difference in this work and that of classical geostatistical estimation techniques is how we treat the below detect data to obtain S-KB. In the classical Kriging case, and to calculate $C_x(p,p')$, we assign to each below detect a value equal to the truncated Gaussian mean as explained earlier. On the other hand in the BME approach we are able to rigorously account for the measurement uncertainty associated with any below detect by selecting a PDF $f_s(x_{\text{soft}}$ that takes the full shape of the Gaussian distribution of PCE concentrations truncated above the detection limit (plain line, figure 1). It follows that the site-specific knowledge consists of the hard data, X_{hard}, corresponding to values measured above their detection limit, and soft data, X_{soft}, corresponding to non-detects that are described using Gaussian PDFs truncated above their detection limits. The overall knowledge bases considered consist of $G = \{m_x(p), C_x(p,p')\}$ and $S = \{f_s(.), X_{hard}\}$. In this case the BME fundamental set of equations reduces to[19] (see also Supporting Information)

$$f_K(x_k) = A^{-1} \int dx\, f_S(x)\, f_G(x) \quad (5)$$

where $f_G(\mathbf{x})$is the Gaussian PDF for X obtained from the G-KB, x is a realization of X, $f_s(x)$ is the truncated Gaussian PDF of X_{soft} and A is a normalization constant.

In this study we average measurements by the year they were sampled due to the lack of temporal variability between wells within the year (See Supporting Information); thus we model the yearly average of PCE concentrations. General and site-specific knowledge were processed as described above by use of BMElib to obtain BME estimates of log-transformed residual S/TRF $X_y(p)$ across North Carolina for each year of the study period. The BME estimate for a given year is a function of data collected in that year, as well as years prior to and after that year. The estimation error associated with BME estimate $X_y(p)$ is fully characterized by the BME posterior PDF. The expected value and corresponding estimation error variance of the corresponding PCE concentration estimate at that estimation point is obtained by adding the global geographical trend $m_z(s)$, and back log-transforming the BME posterior PDF for $X_y(p)$.

This results in BME maps showing the space/time distribution of yearly PCE concentration across North Carolina.

Cross-Validation based on Observed and on Simulated Data

We use a cross validation analysis based on detectable observed PCE concentrations to compare three PCE estimation methods: (SK) Simple Kriging with a constant mean trend and below detect data hardened to the truncated Gaussian mean, (BME) BME with a constant mean trend and below detect treated as the truncated Gaussian PDF, and (LUR/BME) BME with our LUR mean trend and below detect treated as the truncated Gaussian PDF. The cross validation analysis consists in removing each detectable observed log-PCE value Z_j, in turn from the data, and using a given estimation method (k) to calculate its estimate $Z_j^{*(k)}$ based on the remaining data. The mean square error (MSE) calculated as

$$MSE^{(k)} = \frac{1}{n}\sum_{j=1}^{n}\left(Z_j^{*(k)} - Z_j\right)^2 \quad (6)$$

where n is the number of data points, provides a measure of the overall estimation error for method (k). A limitation of cross validation based on detectable observed values is that we do not quantify the estimation error for PCE concentrations that are below the detection limit. To address this issue and quantify the effects of below detect data treatment we also perform cross-validation based on simulated values. Using the covariance model parameters for the constant mean trend Kriging and BME cases, we simulate data at all of the same space/time locations as the real data using the Choleski decomposition method. Then we censor data to simulate below detect observations, where we randomly assign each simulated observation a detection limit value between log(0.1) and log(10), and if the respective simulated observation is less than the detection limit, we treat it as a below detect datum. The random assignment of detection limits is designed to emulate the multiple detection limits that occur when combining data sources. We compare five methods of below detect data treatment with MSE, Pearson's r, Spearman's rho, and Kendall's τ, where the below detect is: (a) assigned to zero, (b) assigned to ½ the detection limit, (c) assigned to the detection limit, (d) treated as a Gaussian PDF with mean and variance equal to the truncated Gaussian mean and variance, and (e) treated as a truncated Gaussian PDF truncated above the detection limit. Methods (a–d) are forms of Kriging estimators, whereas (e) is a novel approach only capable in the BME framework.

Results

GIS Data Integration and Geocoding

Geocoding the private well water resulted in a 4.7 times increase in the amount of PCE data space/time locations from 1999–2010. It also provided a large spatial range of data samples because the EDD monitoring data is spatially clustered, while the private wells are dispersed across the state (Figure S1).

Flow Accumulated Contaminant Source Land Use Regression Model

Contaminant source LUR coefficients and statistics were calculated at regular intervals for the decay range in univariatemodels (Eq. 3). In the univariate case, dry cleaning sites with variable C_{0j} explained the most variability in log-PCE concentrations with the r^2 reaching a maximum of 0.43 with a short decay range of 0.67 km(Table 1, Figure 2), while also having a positive β_1 coefficient. There were no significant bivariate models. We then calculated the flow accumulated model based on the exponential decay from dry cleaners. The model r^2 did not change, but the parameters were significant, so we keep the flow accumulated contaminant source model to favor a more physically meaningful model. Then we calculate the generalized least squared parameters. The intercept changed from −6.47 to −6.86 and the slope changed from 1.02 to 0.75.

Space/Time Covariance Model

The random field $X(p)$ represents PCE concentration with heterogeneity assumed to be removed by the flow routed contaminant source model, thus the homogenous and stationary covariance of $X(p)$ between points $p' = (s',t')$ and can be modeled as being only a function of the spatial lag $r = \|s - s'\|$ and the temporal lag $\tau = |t\text{-}t'|$. Using a numerical algorithm we developed to handle data unevenly distributed over space and time, we calculate experimental covariance values for $X(p)$ by finding pairs (p,p') of measurement events that are separated by various values of r in distance and τ in time. We then used the covariance

experimental values to fit the nonseparable space/time covariance model with a least-squared approach (Figure S3)

$$C_X(r,\tau) = c_1 \exp\left(-\frac{3r}{a_{r_1}}\right)\exp\left(-\frac{3\tau}{a_{\tau_1}}\right) + c_2 \exp\left(-\frac{3r}{a_{r_2}}\right)\exp\left(-\frac{3\tau}{a_{\tau_2}}\right) \quad (7)$$

where $c_1 = 4.09\left(\frac{\mu g}{L}\right)^2$, $a_{r_1} = 5\ km$, $a_{\tau_1} = 7\ years$, $c_2 = 22.33\left(\frac{\mu g}{L}\right)^2$, $a_{r_2} = 0.08\ km$, and $a_{t_2} = 21\ years$.

The least-squared approach for fitting our covariance model produced results that are consistent with PCE transport. The first covariance component explains variability in PCE at long spatial ranges (5 Km) and long temporal ranges (7 years). This component, which makes up approximately 15 % of the total variance, explains persistent plumes associated with PCE.[14] The long spatial component is possibly explained by transport via the air or surface water. The second and major component, which explains approximately 85 % of the total variance, contains a spatial range of 0.60 km and a temporal range of 21 years which is consistent with plume persistence and small scale variability.

Comparison of estimation methods

The cross validation MSE of the SK, BME and LUR/BME methods are, 19.1, 13.1, and 12.1(μg/L)^2, respectively. The combined LUR/BME approach is the best performing method as it reduces the MSE by 7.5% compared to BME which is the second best method. Maps obtained for BME (parts B and D) and LUR/BME (C,E, and G) methods are shown in Figure 3.

Data Simulation Study

The simulated data used leave one out cross validation statistics to compare various approaches to treat the below detects. We find that the BME method (e) decreases the MSE and increases Pearson's r, Spearman's ρ, and Kendall's τ when compared to Kriging methods (a–d) (Table 2).

Discussion

We present an epidemiological concentration estimation framework that can take advantage of all available data including non-georeferenced, primary and secondary data. There are many sources of publicly available datasets for contaminants, but they vary in quality and data reported. For poorly characterized contaminants such as PCE any available dataset should not be overlooked. For PCE, we had two datasets with known latitude and longitude locations and one with only addresses. The 4.7 times increase in georeferenced data points we obtained from our multistage geocoding process was significant because without it we do not obtain meaningful results in the land use regression modeling, nor do the Kriging or BME methods produce informative maps.

North Carolina ranks 4th in total population of self-serviced groundwater use in the US,[24] which potentially contributes to our availability of private well data. We believe that the states within the top 5 of either total population or percent of total population self-serviced groundwater use would benefit from compiling data in a GIS environment, and if necessary, geocoding the well addresses. This would allow them to better protect the health of a high risk group in private well owners. These states are California, Michigan, Pennsylvania, Texas, Maine, Alaska, New Hampshire, Montana, and Wisconsin.

The flow accumulated contaminant source LUR model provides useful information for epidemiological studies in three ways. The LUR identified dry cleaners as important contributors to PCE contamination; moreover, exponential decay range for univariate models corresponds to an isotropic range of influence. We found that dry cleaners sites can influence PCE in groundwater up to 0.67 Km. Second, this method of LUR is applicable to hundreds of anthropogenic contaminants and also potentially useful in identifying exposed populations. Lastly, the LUR model provides the BME method with an informed mean trend leading to more accurate estimates of concentration. Integrating the LUR model into BME as a mean trend directly accounts for spatial auto-correlation in the error leftover from the regression model, which if used as an estimation tool on its own and not accounted for can lead to biased results.[25]

The LUR/BME model accounts for both contaminant sources and for hydrological flow direction and accumulation. It is clear in figure 3F and 3G that the contaminant plume concentration is higher in the direction of decreased elevation gradient. The up gradient side of the source still has concentration due to the isotropic nature of the exponential decay model, but the flow accumulation is responsible for lower values. One can also see from figure 3B–3E that the BME/LUR model leads to a better characterization of concentration close to and far away from dry cleaners. The model is effective at estimating values near dry cleaners, which are generally urbanized, and at estimating far away from dry cleaners, in rural areas where private wells are used. We are confident that the values between DSCA EDD and private well data are comparable since DSCA monitoring wells are sampled at depths in part to reduce the risk of citizens near contaminated sites, which translates to monitoring at depths with human water use availability[26].

The effects of the LUR model are both visible and quantified with lower MSE, however, we recommend that it be used only in the context of a mean trend for a geostatistical model, and not as a standalone model. The results from our maps can be easily incorporated as estimates for concentration for epidemiology studies. Future research investigating associations between PCE and its potential health outcomes can incorporate concentration estimates based upon our methods since it can estimate at any unmonitored location while directly incorporating monitoring data and meaningful mean trends from secondary information.

The results for the LUR/BME model are limited by the available data for the contaminant sources. Since monitoring data were available at dry cleaners, we had reasonable values for a maximum contaminant value at the site. Since this information was not available for the other contaminant sources, the dry cleaners were more informative than any other source. If monitoring data were available at RCRA or CERCLA sites, then we might obtain a result that includes a RCRA or CERCLA term.

Our cross-validation results based on simulated data (Table 2) indicate which treatment of the below detect results in a better estimation of PCE. These results are significant because with contaminants like PCE where the majority of samples result in below detects there is the need to quantify the effectiveness of a model at estimating all values including values that would result in a below detect. Our results show that rather than replacing a below detect with an arbitrary value that is either 0, ½ or 1 times the detection limit (methods a–c), it is better to replace it with a value that accounts for the PCE population distribution truncated above the detection limit (methods d–e), as evidenced by the reduction of the MSE listed in Table 2. Furthermore, our study is the first to find that the best approach is to use a BME approach which rigorously accounts for the full (non-Gaussian) distribution of the below detect (method e) rather than using any of other, Kriging-based, approaches (methods a–d). This is demonstrated by the fact that the cross validation statistics for method (e) listed in Table 2 are consistently better than other statistics listed on that table. For example, the

classical Kriging approach of replacing below detects with zeros leads to a MSE of 7.67($\mu g/L$)^2, while our rigorous non-Gaussian BME approach reduces the MSE to 5.18 ($\mu g/L$)^2. This significant 32% reduction in MSE is accompanied with a consistent increase in Pearson's r, Spearman's rho and Kendall's τ, indicating a consistent reduction in the misclassification of concentration values.

In untreated drinking groundwater epidemiological studies often assume the concentration to equal the exposure. Our results about contaminant concentrations are not to be confused with individual contaminant exposure as cautioned by Jarup,[1] but instead interpreted as an estimation of concentration that one could be exposed to, or as a proxy for exposure. Other factors will determine the concentration an individual is exposed to, which can be determined with GIS techniques as well.[2] Nonetheless, estimates of concentration obtained from our integrated method are useful in identifying high risk populations and in epidemiological studies across large areas where sparse data presents challenges for identifying exposure.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank John Powers and Al Chapman, NC Dry Cleaning Solvent Cleanup Act Program, for providing tetrachloroethylene data. The Development of these maps was supported by the UNC Superfund Research Program-Research Translation Core (P42-ES005948), with funding from the National Institute of Environmental Health Sciences.

References

1. Jarup L. Health and Environmental Information Systems for Exposure and Disease Mapping, and Risk Assessment. Environ. Health Perspect. 2004; 112(9):995–997. [PubMed: 15198919]
2. Nuckols JR, Ward MH, Jarup L. Using Geographic Information Systems for Exposure Assessment in Environmental Epidemiology Studies. Environ Health Perspect. 2004; 112(9):1007–1015. [PubMed: 15198921]
3. Weis BK, Balshaw D, Barr JR, Brown D, Ellisman M, Lioy P, Omenn G, Potter JD, Smith MT, Sohn L, Suk WA, Sumner S, Swenberg J, Walt DR, Watkins S, Thompson C, Wilson SH. Personalized Exposure Assessment: Promising Approaches for Human Environmental Health Research. Environ. Health Perspect. 2005; 113:840–848. [PubMed: 16002370]
4. Eckhardt DAV, Stackelberg PE. Relation of Ground-Water Quality to Land Use on Long Island, New York. Ground Water. 1995; 33(6):1019–1033.
5. Kolpin DW. Agricultural Chemicals in Groundwater of the Midwestern United States: Relations to Land Use. J Environ Qual. 1997; 26:1025–1037.
6. McLay CDA, Dragsten R, Sparling G, Selvarajah N. Predicting groundwater nitrate concentrations in a region of mixed agricultural land use: a comparison of three approaches. Environ. Pollut. 2001; 115:191–204. [PubMed: 11706792]
7. Su JG, Jerrett M, Beckerman B. A distance-decay variable selection strategy for land use regression modeling of ambient air pollution exposures. Sci Total Environ. 2009; 407:3890–3898. [PubMed: 19304313]
8. Akita Y, Carter G, Serre ML. Spatiotemporal Nonattainment Assessment of Surface Water Tetrachloroethylene in New Jersey. J Environ Qual. 2007; 36:508–520. [PubMed: 17332255]
9. Puangthongthub S, Wangwongwatana S, Kamens R, Serre ML. Modeling the Space/Time Distribution of Particulate Matter in Thailand and Optimizing Its Monitoring Network. Atmos. Environ. 2007; 41:7788–7805.
10. Bellander T, Berglind N, Gustavsson P, Jonson T, Nyberg F, Pershagen G, Jarup L. Using Geographic Information Systems to Assess Individual Historical Exposure to Air Pollution from

Traffic and House Heating in Stockholm. Environ. Health Perspect. 2001; 109(6):633–639. [PubMed: 11445519]

11. Moran MJ, Zogorski JS, Squillance PJ. Chlorinated Solvents in Groundwater of the United States. Environ. Sci. Technol. 2007; 41:74–81. [PubMed: 17265929]
12. Agency for Toxic Substance Disease Registry. Toxicological Profile for Tetrachloroethylene. Atlanta, GA: 1997.
13. National Research Council. Review of the Environmental Protection Agency's Draft IRIS Assessment of Tetrachloroethylene. National Academies Press; Washington, DC: 2010.
14. Dry Cleaning Solvent and Cleanup Act Program. [(accessed June 1, 2010)] Sections and Programs of the North Carolina Division of Waste Management. http://portal.ncdenr.org/web/wm/dsca
15. Safe Drinking Water/Private Wells. North Carolina General Assembly. House Bill 2873. Session Law 2006-202; http://www.ncga.state.nc.us/sessions/2005/bills/house/pdf/h2873v6.pdf
16. United States Geological Survey. [(accessed April 1, 2010)] National Water Information System. http://nwis.waterdata.usgs.gov
17. NCOnemap. [(accessed June 1, 2009)] Geographic Data Serving A Statewide Community. http://www.nconemap.com/
18. Schwanghart W, Kuhn NJ. TopoToolbox : a set of Matlab functions for topographic analysis. Environ. Modell. Softw. 2010; 25:770–781.
19. Christakos G. A Bayesian/Maximum-Entropy View To The Spatial Estimation Problem. Math. Geosci. 1990; 22(7):763–776.
20. Christakos, G. Modern Spatiotemporal Geostatistics. Oxford University Press; New York: 2000.
21. De Nazelle A, Arunachalam S, Serre ML. Bayesian Maximum Entropy Integration of Ozone Observations and Model Predictions: An Application for Attainment Demonstration in North Carolina. Environ. Sci. Technol. 2010; 44(15):5707–5713. [PubMed: 20590110]
22. Serre ML, Christakos G. Modern geostatistics: Computational BME analysis in the light of uncertain physical knowledge – the Equus Beds study. Stoch Environ Res Risk Assess. 1999; 13(1):1–26.
23. Christakos, G.; Bogaert, P.; Serre, ML. Temporal GIS: Advanced Functions for field-based Applications. Springer; New York: 2002.
24. Kenny JF, Barber NL, Hutson SS, Linsey KS, Lovelace JK, Maupin MA. Estimated use of water in the United States in 2005. US Geologic Survey Circular 1344. 2009; 52:20.
25. Qian SS, Reckow KH, Zhai J, McMahon G. Nonlinear regression modeling of nutrient loads in streams: A Bayesian approach. Water Resour. Res. 2005; 41:W07012. doi: 10.1029/2005WR003986.
26. [(Accessed December 10, 2011)] Rules and Criteria for the Administration of the Dry-Cleaning Solvent Cleanup Fund. 15A NCAC 02S; http://portal.ncdenr.org/web/dm/dsca

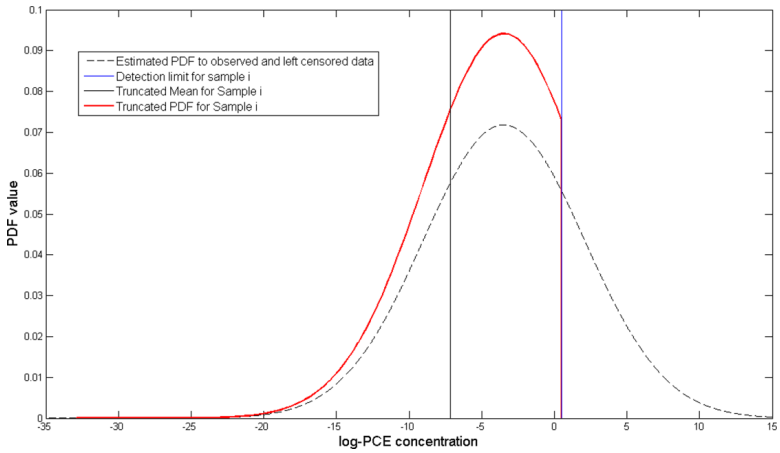

Figure 1.

PDF of log-PCE with mean and variance estimated from observed and left censored data (see Supporting Information), showing a sample detection limit and corresponding truncated Gaussian mean

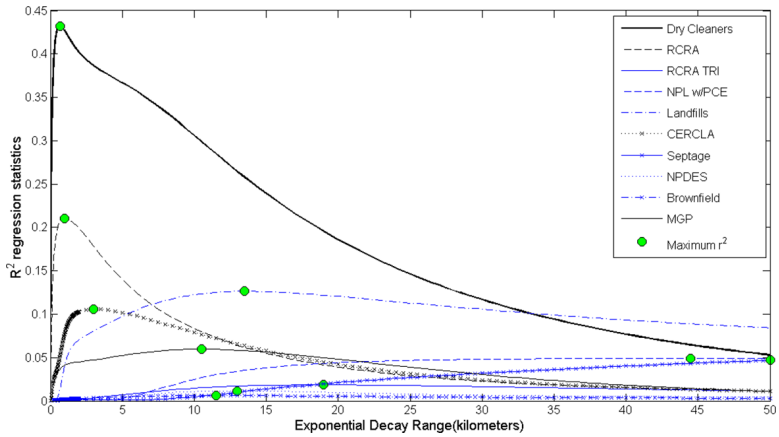

Figure 2.
r^2 regression statistics as a function of the exponential decay range

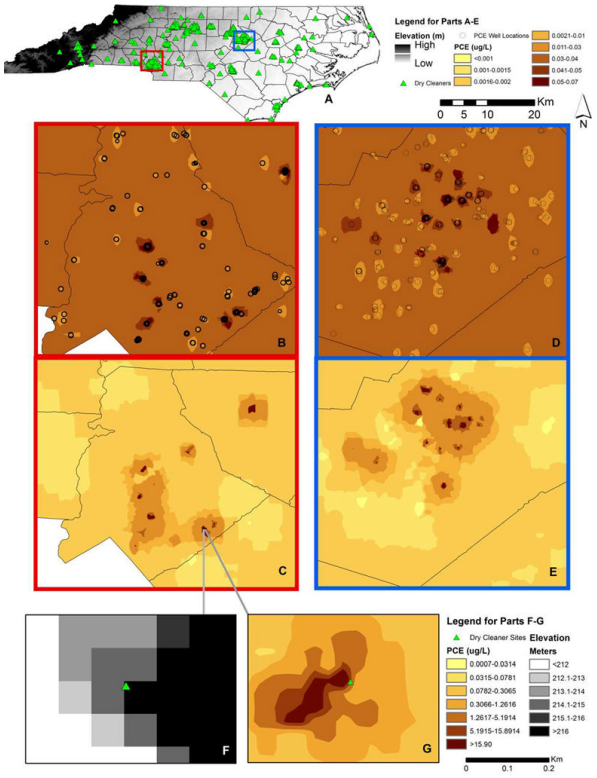

Figure 3.

(A) North Carolina elevation and Dry Cleaner locations. The red extent rectangle corresponds to parts B and C. The blue extent rectangle corresponds with parts D and E. (B) BME estimate with below detects treated as a truncated Gaussian PDF for a selected county. (C) LUR and BME estimate with below detects treated as a truncated Gaussian PDF and a land use regression mean trend based on dry cleaners for a selected county (same as B) (D) BME estimate at another selected county. (E) LUR/BME estimate at another selected county (same as D). (F) Elevation for site in part G. (G) LUR/BME estimate at a dry cleaner site. Note how the plume roughly follows the change in elevation.

Table 1

Statistics for univariate land use regression models obtained for the decay range corresponding to the maximum r-squared value.

Contaminant Source Variable	Exponential Decay Range in km	r-squared regression statistic	p-value (F-Stat)	Beta 1 (95% CI)
Dry Cleaning	0.67	0.43	<0.0001 (4112.1)	1.01 (0.98–1.04))
RCRA	0.97	0.21	<0.0001 (14398)	2.04 (1.93–2.15)
RCRA TRI	19.0	0.0186	<0.0001 (102)	3.91 (3.15–4.67)
CERCLA	3.0	0.11	<0.0001 (639.9)	2.34(2.15–2.52)
NPL w/PCE	44.5	0.05	<0.0001(277.9)	1.12 (0.99–1.26)
NPDES	13.0	0.01	<0.0001 (63.95)	037 (0.28–0.46)
Landfill	13.5	0.05	<0.0001 (265.4)	0.55(0.49– 0.61)
Brownfield	11.5	0.01	<0.0001 (33.9)	0.14(0.09–0.19)
Manufacturing Gas Plant	10.5	0.06	<0.0001 (342.8)	4.50(4.02–4.98)
Septage	50.0	0.05	<0.0001 (265.4)	0.55 (0.49–0.61)

Table 2

Validation statistics for the simulation study comparing methods of below detect data treatment

Below Detect Method	MSE($\mu g/L$)^2	Pearson's r	Spearman's rho	Kendall's tau
Hardended to Zero	7.67	0.67	0.67	0.48
Hardened to ½ Detection Limit	7.58	0.61	0.54	0.39
Hardened to Detection Limit	6.65	0.64	0.60	0.43
Truncated Gaussian Approximation with Kriging	5.34	0.70	0.68	0.39
Truncated Gaussian with BME	5.18	0.71	0.69	0.51